

A FAST CLUSTERING – BASED HIGH – DIMENSIONAL DATA BY USING TEXT CLASSIFICATION

M. Amudha¹

T. Bhuvaneshwari¹

¹ Final Year Students, Department of Computer Science and Engineering, E.G.S. Pillay Engineering College, Nagapattinam, Tamil Nadu, India.

M. Rajakumaran²

P. Anandraj²

T. Ganesan²

² Assistant Professor, Department of Computer Science and Engineering, E.G.S. Pillay Engineering College, Nagapattinam, Tamil Nadu, India.

ARTICLE INFO

Article History:

Received: 08 Mar 2016;

Received in revised form:

14 Mar 2016;

Accepted: 14 Mar 2016;

Published online: 31 Mar 2016.

Key words:

Text Mining,

Text Classification,

Information Filtering

ABSTRACT

Most existing popular text segregation methods have adopted term-based approaches. It classifies terms into categories and updates term weights based on their specificity and their distributions in patterns. The field of text mining seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns. The discovery of relevant features in real-world data for describing user information needs or preferences is a new challenge in text mining. Relevance of a feature indicates that the features is always necessary for an optimal subset, it cannot be removed without affecting the original conditional class distribution. In this paper, an adaptive method for relevance feature discovery is discussed, to find useful features available in a feedback set, including both positive and negative documents, for describing what users need. Thus, this paper discusses the methods for relevance feature discovery using the simulated annealing approximation and genetic algorithm, a population of candidate solutions to an optimization problem toward better solutions.

Copyright © 2016 IJASRD. This is an open access article distributed under the Creative Common Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

INTRODUCTION

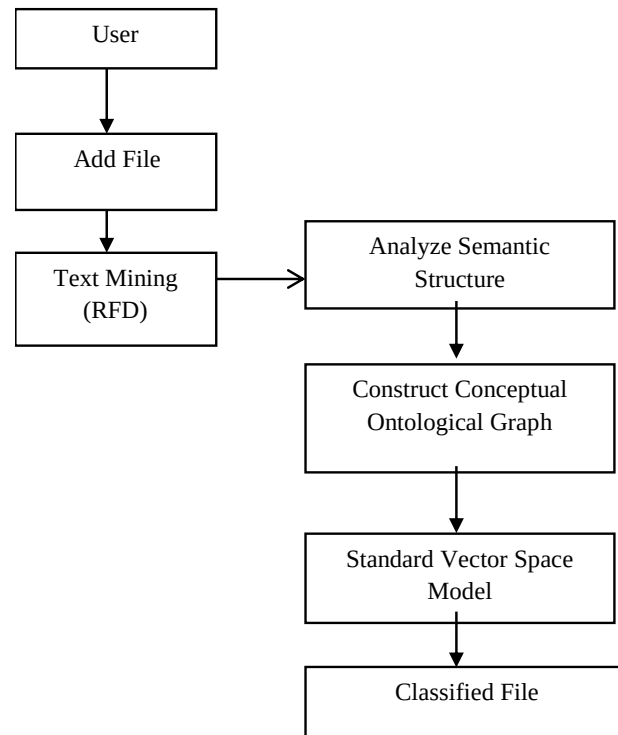
Text mining is a burgeoning new field that attempts to glean meaningful information from natural language text. It may be loosely characterized as the process of analyzing text to extract information that is useful for particular purposes. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. Nevertheless, in modern culture, text is the most common vehicle for the formal exchange of information. The field of text mining usually deals with texts whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling—even if success is only partial. The phrase “text mining” is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information [Sebastian, 2002]. In discussing a topic that lacks a generally accepted definition in a practical Handbook such as this, I have chosen to cast the net widely and take a liberal viewpoint of what should be included, rather than attempting a clear-cut characterization that will inevitably restrict the scope of what is covered.

1.1 Data Mining and Text Mining:

Just as data mining can be loosely described as looking for patterns in data, text mining is about looking for patterns in text. However, the superficial similarity between the two conceals real differences. Data mining can be more fully characterized as the extraction of implicit, previously unknown, and potentially useful information from data [Witten and Frank, 2000]. The information is implicit in the input data: it is hidden, unknown, and could hardly be extracted without recourse to automatic techniques of data mining. With text mining, however, the information to be extracted is clearly and explicitly stated in the text. It’s not hidden at all—most authors go to great pains to make sure that they express themselves clearly and unambiguously—and, from a human point of view, the only sense in which it is “previously unknown” is that human resource restrictions make it infeasible for people to read the text themselves. The problem, of course, is that the information is not couched in a manner that is amenable to automatic processing. Text mining strives to bring it out of the text in a form that is suitable for consumption by computers directly, with no need for a human intermediary though there is a clear difference philosophically, from the computer’s point of view the problems are quite similar. Text is just as opaque as raw data when it comes to extracting information—probably more so. Another requirement that is common to both data and text mining is that the information extracted should be “potentially useful.” In one sense, this means *actionable*—capable of providing a basis for actions to be taken automatically. In the case of data mining, this notion can be expressed in a relatively domain-independent way: actionable patterns are ones that allow non-trivial predictions to be made on new data from the same source. Performance can be measured by counting successes and failures, statistical techniques can be applied to compare different data mining methods on the same problem, and so on. However, in much text mining situations it is far harder to characterize what “actionable” means in a way that is independent of the particular domain at hand. This makes it difficult to find fair and objective measures of success. In many data mining applications, “potentially useful” is

given a different interpretation: the key for success is that the information extracted must be *comprehensible* in that it helps to explain the data. This is necessary whenever the result is intended for human consumption rather than (or as well as) a basis for automatic action. This criterion is less applicable to text mining because, unlike data mining, the input itself is comprehensible. Text mining with comprehensible output is tantamount to summarizing salient features from a large body of text, which is a subfield in its own right: text summarization.

Fig 1: Architectural Design



1.2 Natural Language Analysis Techniques

In this section examine in some detail several of the more common techniques for natural language analysis, i.e. for translating natural language utterances into a unique internal representation. Virtually all natural language analysis systems can be classified into one of the following categories:

- Pattern matching (e.g. ELIZA [72], PARRY [51])
- Syntactically-driven parsing (e.g. ATNs [75])
- Semantic grammars (e.g. LIFER [41], SOPHIE [6])
- Case frame instantiation (e.g. ELI [55])
- Wait and see (e.g. Marcus [48])
- Natural Language Analysis Techniques 7
- Word expert (e.g. Small [68])
- Connectionist (e.g. Small [69])
- Skimming (e.g. FRUMP [26], IPP [63])

The examples provided with each category are the names of language analysis systems following that approach, or the names of builders of such systems. Of these categories, the first four represent the bulk of the language analysis systems already

constructed, and are the only ones we will cover in detail. The reader is encouraged to follow up the references provided for further details of the other methods.

1.3 Pattern Matching

The essence of the pattern matching approach to natural language analysis is to interpret input utterances as a whole, rather than building up their interpretation by combining the structure and meaning of words or other lower-level constituents. The approach is thus wholistic rather than constructive. With this approach, the interpretations are obtained by matching patterns of words against the input utterance. Associated with each pattern is an interpretation, so that the derived interpretation is the one attached to the pattern that matched.

SYSTEM ANALYSIS

2.1 Existing System:

Pattern taxonomy mining (PTM) models have been proposed in which, mining closed sequential patterns in text paragraphs and deploying them over a term space to weight useful features. Concept-based model (CBM) has also been proposed to discover concepts by using natural language processing (NLP) techniques. However, fewer significant improvements are made compared with the best term-based method because how to effectively integrate patterns in both relevant and irrelevant documents is still an open problem.

2.2 Proposed System:

Relevance is a big research issue, for Web search, which discusses a documents relevance to a user or a query. The efficient way of feature selection for relevance is based on a feature weighting function. A feature weighting function indicates the degree of information represented by the feature occurrences in a document and reflects the relevance of the feature. Text features can be simple structures (words), complex linguistic structures or statistical structures.

SYSTEM IMPLEMENTATION

3.1 Module Split Up

- Text Mining
- Text Feature Extraction
- Text Classification
- Summarization
- Semantic Text File

3.2 Modules Description

3.2.1 Text Mining

Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them. Hence, you can analyze words, clusters of words used in documents, etc., or you could

analyze documents and determine similarities between them or how they are related to other variables of interest in the data mining project. In the most general terms, text mining will “turn text into numbers” (meaningful indices), which can then be incorporated in other analyses such as predictive data mining projects, the application of unsupervised learning methods (clustering), etc.

3.2.2 Text Feature Extraction

Feature selection: reduce number of feature and selection subset which best represent data

Feature extraction: select subset but into new space

During training, a feature extractor is used to convert each input value to a feature set. These feature sets, which capture the basic information about each input that should be used to classify it, are discussed in the next section. Pairs of feature sets and labels are fed into the machine learning algorithm to generate a model.

During prediction, the same feature extractor is used to convert unseen inputs to feature sets. These feature sets are then fed into the model, which generates predicted labels.

3.2.3 Text Classification

Classification tasks can be divided into three sorts:

- Supervised Document Classification
- Unsupervised Document Classification
- Semi-Supervised Document Classification

Supervised document classification where some external mechanism (such as human feedback) provides information on the correct classification for documents, unsupervised document classification (also known as document clustering), where the classification must be done entirely without reference to external information, and semi-supervised document classification, where parts of the documents are labeled by the external mechanism.

3.2.4 Summarization

Effective summarizing requires an explicit and detailed analysis of context factors. The features of the text to be summarized crucially determine the way a summary can be obtained. Location refers to the position in text, paragraph or any other particular section in the sense that they contain the target sentences to be included in the summary.

Similarity occurs for example, when two words share a common stem, i.e. whose form is similar. This can be extended for phrases or paragraphs. Similarity can be calculated by vocabulary overlap or with linguistic techniques.

3.2.5 Semantic Text File

“Semantic Analysis” refers to a formal analysis of meaning, and “computational” refer to approaches that in principle support effective implementation. A text consists of various worlds in which different actors do, form or say various things in combination with other actors. There must be a theoretical conception of the text: this must describe both the textual organization of the things that are said and the structural organization of the

thought-processes of the people who say them. It implies the use of a tool derived from this theoretical conception and which rigorously excludes the subjectivity of the investigator – at least until the analysis is finished.

LITERATURE SURVEY

Effective Pattern Discovery for Text Mining, Author Ning Zhong, Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis.

A Comparison of Document Clustering Techniques, Author Vipin Kumar, This paper presents the results of an experimental study of some common document clustering techniques. In particular, we compare the two main approaches to document clustering, agglomerative hierarchical clustering and K-means.

The Fuzzy C-Means Clustering Algorithm, Author William Full, In this paper transmits a FORTRAN-IV coding of the fuzzy c-means (FCM) clustering program. The FCM program is applicable to a wide variety of geostatistical data analysis problems. This program generates fuzzy partitions and prototypes for any set of numerical data.

RESULT

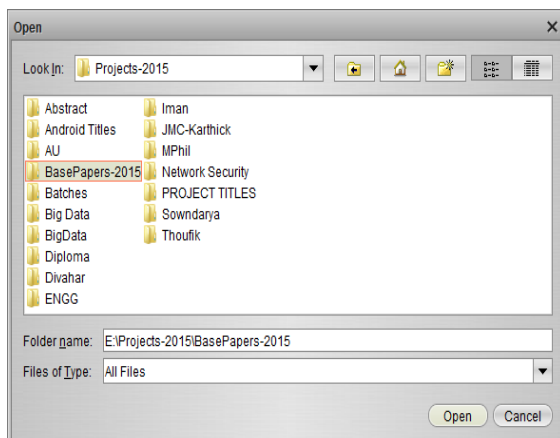


Fig 2: Select Folder



Fig 3: Extract Text & Images from PDF



Fig 4: Clustering & Summarize text

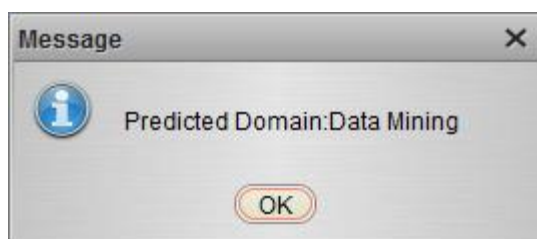


Fig 5: Predict Domain

CONCLUSION

In this project, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves 1) removing irrelevant features, 2) constructing a minimum spanning tree from relative ones, and 3) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. We have compared the performance of the proposed algorithm with those of the five well-known feature selection algorithms FCBF, Relief, CFS, Consist, and FOCUS-SF on the 35 publicly available image, microarray, and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record. Generally, the proposed algorithm obtained the best proportion of selected features, the best runtime, and the best classification accuracy for Naive Bayes, C4.5, and RIPPER, and the second best classification.

FUTURE ENHANCEMENT

In this research work, an effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. The experimental results show that the proposed model outperforms not only other pure data mining-based methods and the concept based model, but also term-based state-of-the-art models, such as BM25 and SVM-based models.

REFERENCES

- [1] Aas. K and Eikvil. L, "Text Categorisation: A Survey," Technical Report Raport NR 941, Norwegian Computing Center, 1999.
- [2] Agrawal. R and Srikant. R, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 478-499, 1994.
- [3] Ahonen. H, Heinonen. O, Klemettinen. M, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.
- [4] Baeza-Yates. R and Ribeiro-Neto. B, Modern Information Retrieval. Addison Wesley, 1999.
- [5] Cancedda. N, Cesa-Bianchi. N, Conconi. A, and Gentile. C, "Kernel Methods for Document Filtering," TREC, 2002.
- [6] Cancedda. N, Gaussier. E, Goutte. C, and J.-M. Renders, "Word- Sequence Kernels," J. Machine Learning Research, vol. 3, pp. 1059- 1082, 2003.
- [7] Caropreso. M, Matwin. S, and Sebastiani. F, "Statistical Phrases in Automated Text Categorization," Technical Report IEI-B4-07- 2000, Istituto di Elaborazione dell'Informazione, 2000.
- [8] Cortes. C and Vapnik. V, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

- [9] Dumais. S, “Improving the Retrieval of Information from External Sources,” Behavior Research Methods, Instruments, and Computers, vol. 23, no. 2, pp. 229-236, 1991.
- [10] Han. J and Chang. C, “Data Mining for Web Intelligence,” Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.
- [11] Han. J, Pei. j, and Yin. Y, “Mining Frequent Patterns without Candidate Generation,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’00), pp. 1-12, 2000.
- [12] Huang. Y and Lin. S, “Mining Sequential Patterns Using Graph Search Techniques,” Proc. 27th Ann. Int’l Computer Software and Applications Conf., pp. 4-9, 2003.
- [13] Jindal. N and Liu. B, “Identifying Comparative Sentences in Text Documents,” Proc. 29th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’06), pp. 244-251, 2006.
- [14] Joachims. T, “A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization,” Proc. 14th Int’l Conf. Machine Learning (ICML ’97), pp. 143-151, 1997.
- [15] Joachims. T, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” Proc. European Conf. Machine Learning (ICML ’98), pp. 137-142, 1998.
- [16] Joachims. T, “Transductive Inference for Text Classification Using Support Vector Machines,” Proc. 16th Int’l Conf. Machine Learning (ICML ’99), pp. 200-209, 1999.
- [17] Lam. W, Ruiz. M, and Srinivasan. P, “Automatic Text Categorization and Its Application to Text Retrieval,” IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.
- [18] Lewis. D, “An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task,” Proc. 15th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’92), pp. 37-50, 1992.
- [19] Lewis. D, “Feature Selection and Feature Extraction for Text Categorization,” Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.
- [20] Lewis. D, “Evaluating and Optimizing Automatic Text Classification Systems,” Proc. 18th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’95), pp. 246-254, 1995.